APPENDIX

A. Proof of Theorem 1

Let us rewrite the risks and generalization error under the DNN setup. Let $(X,Y)\sim P_{X,Y}$ be a pair of test data sample. At each layer l, the internal representation T_l of a test data feature X is conditionally independent of W_{l+1}^L given W_1^l. For any $\mathbf{W}\in\mathcal{W}$, let the loss function be rewritten as $\ell(\mathbf{W},X,Y)=\ell(g_{\mathbf{W}_L}\circ g_{\mathbf{W}_{L-1}}\circ\cdots\circ g_{\mathbf{W}_1}(X),Y)$. The expected population risk over all possible $\mathbf{W}$ is given by

$$\mathbb{E}_W[\mathcal{L}_{\mathsf{P}}(\mathbf{W},P_{X,Y})]=\mathbb{E}[\mathbb{E}[\ell(g_{\mathbf{W}_L}\circ g_{\mathbf{W}_{L-1}}\circ\cdots\circ g_{\mathbf{W}_{l+1}}(T_l),Y)|\mathbf{W}_1^l]]$$

where $l\in[L]$ and given $\mathbf{W}_1^l$, (T_l,Y) are independent of $\mathbf{W}_{l+1}^L$.

Denote the overall feature mapping function as $f_{\mathbf{W}}\triangleq g_{\mathbf{W}_L}\circ g_{\mathbf{W}_{L-1}}\circ\cdots\circ g_{\mathbf{W}_1}$. Similarly, for any $l\in[L]$, the expected empirical risk can also be rewritten as

$$\mathbb{E}[\mathcal{L}_{\mathsf{E}}(\mathbf{W},D_n)]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\mathbb{E}[\ell(g_{\mathbf{W}_L}\circ g_{\mathbf{W}_{L-1}}\circ\cdots\circ g_{\mathbf{W}_{l+1}}(T_{l,i}),Y_i)|\mathbf{W}_1^l]\big].$$

For notational simplicity, let $g_{\mathbf{W}_k^j}:=g_{\mathbf{W}_k}\circ g_{\mathbf{W}_{k-1}}\circ\cdots\circ g_{\mathbf{W}_j}$ for any $k<j$ and $k,j\in\mathbb{N}$. Then the expected generalization error can be rewritten as

$$\mathsf{gen}(P_{\mathbf{W}|D_n},P_{X,Y})=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\mathbb{E}[\ell(g_{\mathbf{W}_{l+1}^L}(T_l),Y)|\mathbf{W}_1^l]-\mathbb{E}[\ell(g_{\mathbf{W}_{l+1}^L}(T_{l,i}),Y_i)|\mathbf{W}_1^l]\Big].\tag{2}$$

If the loss function $\ell(\mathbf{w},X,Y)$ is σ-sub-Gaussian under $P_{X,Y}$ for all $\mathbf{w}\in\mathcal{W}$, we also have for any $l\in[0:L]$, $\ell(g_{\mathbf{w}_{l+1}^L}(T_l),Y)$ is σ-sub-Gaussian under $P_{T_l,Y|\mathbf{W}=\mathbf{w}}$ for all $\mathbf{w}\in\mathcal{W}$. From Donsker-Varadhan representation, we have for any $\lambda\in\mathbb{R}$,

$$\begin{aligned}&\mathsf{D}_{\mathsf{KL}}(P_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}\otimes P_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l})\\&\geq\mathbb{E}_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}[\lambda\ell(g_{\mathbf{W}_{l+1}^L}(T_{l,i}),Y_i)]-\log\mathbb{E}_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}\mathbb{E}_{T_l,Y|\mathbf{W}_1^l}[\exp(\lambda\ell(g_{\mathbf{W}_{l+1}^L}(T_l),Y))]\\&\geq\lambda(\mathbb{E}_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}[\lambda\ell(g_{\mathbf{W}_{l+1}^L}(T_{l,i}),Y_i)]-\mathbb{E}_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}\mathbb{E}_{T_l,Y|\mathbf{W}_1^l}[\ell(g_{\mathbf{W}_{l+1}^L}(T_l),Y)])-\frac{\lambda^2\sigma^2}{2}.\end{aligned}$$

We can decompose $\mathsf{D}_{\mathsf{KL}}(P_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}\otimes P_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})$ as follows

$$\begin{aligned}&\mathsf{D}_{\mathsf{KL}}(P_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}\otimes P_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})\\&=\mathsf{D}_{\mathsf{KL}}(P_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_{l,i},Y_i|\mathbf{W}_1^l}\otimes P_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})+\mathsf{D}_{\mathsf{KL}}(P_{T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})\\&=\mathsf{I}(T_{l,i},Y_i;\mathbf{W}_{l+1}^L|\mathbf{W}_1^l)+\mathsf{D}_{\mathsf{KL}}(P_{T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l}).\end{aligned}\tag{3}$$

Thus, we have

$$\begin{aligned}&\mathsf{I}(T_{l,i},Y_i;\mathbf{W}_{l+1}^L|\mathbf{W}_1^l)+\mathsf{D}_{\mathsf{KL}}(P_{T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})=\mathsf{D}_{\mathsf{KL}}(P_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}\otimes P_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})\\&\geq\lambda\mathbb{E}_{\mathbf{W}_1^l}\left[\mathbb{E}_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}[\ell(g_{\mathbf{W}_{l+1}^L}(T_{l,i}),Y_i)]-\mathbb{E}_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}\mathbb{E}_{T_l,Y|\mathbf{W}_1^l}[\ell(g_{\mathbf{W}_{l+1}^L}(T_l),Y)]\right]-\frac{\lambda^2\sigma^2}{2}.\end{aligned}$$

By optimizing the RHS over $\lambda>0$ and $\lambda\leq 0$, respectively, we finally obtain

$$\begin{aligned}&\Big|\mathbb{E}_{\mathbf{W}_1^l}\mathbb{E}_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}[\ell(g_{\mathbf{W}_{l+1}^L}(T_{l,i}),Y_i)]-\mathbb{E}_{\mathbf{W}_1^l}\mathbb{E}_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}\mathbb{E}_{T_l,Y|\mathbf{W}_1^l}[\ell(g_{\mathbf{W}_{l+1}^L}(T_l),Y)]\Big|\\&\leq\sqrt{2\sigma^2\big(\mathsf{I}(T_{l,i},Y_i;\mathbf{W}_{l+1}^L|\mathbf{W}_1^l)+\mathsf{D}_{\mathsf{KL}}(P_{T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})\big)},\end{aligned}$$

which holds for all $l\in[L]$. Conditioned on $\mathbf{W}_l$, $T_{l,i}$ and T_l are generated by the same process from $T_{l-1,i}$ and T_{l-1}, respectively. By the data-processing inequality, the KL divergence in (3) can be bounded as follows:

$$\begin{aligned}\mathsf{D}_{\mathsf{KL}}(P_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}\otimes P_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})&\leq\mathsf{D}_{\mathsf{KL}}(P_{\mathbf{W}_{l+1}^L,T_{l-1,i},Y_i|\mathbf{W}_1^l}\|P_{T_{l-1},Y|\mathbf{W}_1^l}\otimes P_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})\\&=\mathsf{D}_{\mathsf{KL}}(P_{\mathbf{W}_l^L,T_{l-1,i},Y_i|\mathbf{W}_1^{l-1}}\|P_{T_{l-1},Y|\mathbf{W}_1^{l-1}}\otimes P_{\mathbf{W}_l^L|\mathbf{W}_1^{l-1}}|P_{\mathbf{W}_1^{l-1}})\\&\;\;\vdots\\&\leq\mathsf{D}_{\mathsf{KL}}(P_{X_i,Y_i,\mathbf{W}_1^L}\|P_{X,Y}\otimes P_{\mathbf{W}_1^L})=\mathsf{I}(X_i,Y_i;\mathbf{W}).\end{aligned}$$

Theorem 1 can be thus proved by induction.

B. Proof of Theorem 2

Recall the Kantorovich-Rubinstein duality [36]: for any two probability measures $P, Q \in \mathcal{P}(\mathcal{X})$, $\mathsf{W}_1(P,Q) = \sup_{f\in \mathrm{Lip}_1(\mathcal{X})} \mathbb{E}_P[f] - \mathbb{E}_Q[f]$, where $\mathrm{Lip}_k(\mathcal{X}) = \{f \in \{f : \mathcal{X} \to \mathbb{R}\} : |f(x) - f(y)| \le k\|x-y\|, \forall x, y \in \mathcal{X}\}$, for any $k \in \mathbb{R}_{\ge 0}$

Since $\tilde{\ell}(g_{\mathbf{W}_L} \circ \cdots \circ g_{\mathbf{W}_1}(X), Y)$ is ρ_0-Lipschitz in $(g_{\mathbf{W}_L} \circ \cdots \circ g_{\mathbf{W}_1}(X), Y)$ and $\phi_l(\cdot)$ is ρ_l-Lipschitz, we have for any $\mathbf{w}$,

$$\begin{aligned}|\ell(\mathbf{w},x,y) - \ell(\mathbf{w},x',y')| &= |\tilde{\ell}(g_{\mathbf{w}_L} \circ \cdots \circ g_{\mathbf{w}_1}(x), y) - \tilde{\ell}(g_{\mathbf{w}_L} \circ \cdots \circ g_{\mathbf{w}_1}(x'), y')| \\ &\le \rho_0 \|(g_{\mathbf{w}_L}(g_{\mathbf{w}_1^{L-1}}(x)), y) - (g_{\mathbf{w}_L}(g_{\mathbf{w}_1^{L-1}}(x')), y')\| \\ &\le \rho_0 \sqrt{\left(\rho_L \|\mathbf{w}_L\| \|g_{\mathbf{w}_1^{L-1}}(x) - g_{\mathbf{w}_1^{L-1}}(x')\|\right)^2 + (y-y')^2} \\ &\vdots \\ &\le \bar{\rho}_0(\mathbf{w}) \sqrt{\|x-x'\|^2 + (y-y')^2}\end{aligned}$$

where $\bar{\rho}_0(\mathbf{w}) \coloneqq \rho_0(1 \vee \prod_{j=1}^L \rho_j \|\mathbf{w}_j\|)$. It means $\ell(\mathbf{W}, X, Y)$ is $\bar{\rho}_0(\mathbf{W})$-Lipschitz in (X,Y) for any $\mathbf{W}$. Then we have

$$\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y}) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\ell(\mathbf{W},X,Y) - \ell(\mathbf{W},X_i,Y_i)] \le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\bar{\rho}_0(\mathbf{W})\mathsf{W}_1(P_{X_i,Y_i|\mathbf{W}}, P_{X,Y|\mathbf{W}})\right].$$

For $l = 1, \dots, L$, similarly, we have $\tilde{\ell}(g_{\mathbf{W}_l} \circ \cdots \circ g_{\mathbf{W}_{l+1}}(T_l), Y)$ is $\rho_0(1 \vee \prod_{j=l+1}^L \rho_j \|\mathbf{W}_j\|)$-Lipschitz in (T_l, Y) for all $\mathbf{W}$. Let $\bar{\rho}_l(\mathbf{W}) = \rho_0(1 \vee \prod_{j=l+1}^L \rho_j \|\mathbf{W}_j\|)$. Then from the definition in (2), we have

$$\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y}) \le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\bar{\rho}_l(\mathbf{W})\mathsf{W}_1(P_{T_{l,i},Y_i|\mathbf{W}}, P_{T_l,Y|\mathbf{W}})\right].$$

The proof is completed by taking the minimum over $l = 0, \dots, L$.

C. Proof of Remark 1

Let $\mathsf{diam}(\mathcal{X}) \coloneqq \sup\{\|x-y\| : x, y \in \mathcal{X}\}$. From [35, Theorem 4], Pinsker's and Bretagnolle-Huber inequalities, for any two probability distributions $\mu, \nu \in \mathcal{P}(\mathcal{X})$, we have

$$\mathsf{W}_1(\mu,\nu) \le \mathsf{diam}(\mathcal{X})\mathsf{D}_{\mathsf{TV}}(\mu,\nu) \le \mathsf{diam}(\mathcal{X})\sqrt{\left(\frac{1}{2}\mathsf{D}_{\mathsf{KL}}(\mu\|\nu) \wedge \left(1 - \exp(-\mathsf{D}_{\mathsf{KL}}(\mu\|\nu))\right)\right)}.$$

From Theorem 1 and [37], the generalization error can be bounded as follows:

$$\left|\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})\right| \le \frac{A}{n}\sum_{i=1}^n \mathsf{D}_{\mathsf{TV}}(P_{T_{L,i},Y_i|\mathbf{W}}, P_{T_L,Y|\mathbf{W}}|P_{\mathbf{W}}) \le \mathsf{UB}(L),$$

where $\mathsf{UB}(L) \coloneqq \frac{\sqrt{2}A}{2n}\sum\limits_{i=1}^n \sqrt{\mathsf{D}_{\mathsf{KL}}\left(P_{T_{L,i},Y_i|\mathbf{W}} \middle\| P_{T_L,Y|\mathbf{W}}|P_{\mathbf{W}}\right)}$. It can be observed that the total variation distance based bound is tighter under this condition.

Let $l^* \coloneqq \min_{l=0,\dots,L} \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\bar{\rho}_l(\mathbf{W})\mathsf{W}_1(P_{T_{l,i},Y_i|\mathbf{W}}, P_{T_{l},Y|\mathbf{W}})]$.

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\bar{\rho}_{l^*}(\mathbf{W})\mathsf{W}_1(P_{T_{l^*,i},Y_i|\mathbf{W}}, P_{T_{l^*},Y|\mathbf{W}})] \le \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\bar{\rho}_L(\mathbf{W})\mathsf{W}_1(P_{T_{L,i},Y_i|\mathbf{W}}, P_{T_L,Y|\mathbf{W}})] \\ &= \frac{\rho_0}{n}\sum_{i=1}^n \mathsf{W}_1(P_{T_{L,i},Y_i|\mathbf{W}}, P_{T_L,Y|\mathbf{W}}|P_{\mathbf{W}}) \le \frac{\rho_0 \mathsf{diam}(\mathcal{T}_L \times \mathcal{Y})}{n}\sum_{i=1}^n \mathsf{D}_{\mathsf{TV}}(P_{T_{L,i},Y_i|\mathbf{W}}, P_{T_L,Y|\mathbf{W}}|P_{\mathbf{W}}).\end{aligned}$$

Under our supervised classification setting, $\mathsf{diam}(\mathcal{T}_L \times \mathcal{Y}) = K^2$. Thus, when $\rho_0 K^2 \le A$, 1-Wasserstein distance based bound is even tighter than the one based on the TV-distance.

D. Proofs for the case study of binary Gaussian mixture classification

To simplify some of the notation ahead, the distribution of a Gaussian random variable X with mean μ and variance σ^2 is denoted by $\mathcal{N}_X(\mu,\sigma^2)$. Under the binary Gaussian mixture classification setting, we first know that the prior of $\mathbf{W}_{\otimes L}^\intercal$ is $P_{\mathbf{W}_{\otimes L}^\intercal} = \mathcal{N}_{\mathbf{W}_{\otimes L}}(\mu_0, \frac{\sigma_0^2}{n}\mathbf{I}_{d_0})$. Given any pair of training data sample (X_i, Y_i), we have

$$\mathbf{W}_{\otimes L}^\intercal|(X_i,Y_i) = \frac{1}{n}Y_iX_i + \frac{1}{n}\sum_{j\neq i}^{n} Y_jX_j \sim \mathcal{N}_{\mathbf{W}_{\otimes L}}(\mu_{\mathbf{W}_{\otimes L}|i}, \mathbf{\Sigma}_{\mathbf{W}_{\otimes L}|i}),$$

where $\mu_{\mathbf{W}_{\otimes L}|i} = \frac{1}{n}Y_iX_i + \frac{n-1}{n}\mu_0$ and $\mathbf{\Sigma}_{\mathbf{W}_{\otimes L}|i} = \frac{n-1}{n^2}\sigma_0^2\mathbf{I}_{d_0}$. Then the posterior distribution of (X_i,Y_i) given $\mathbf{W}_{\otimes L}$ is given by

$$\begin{aligned}
P_{X_i,Y_i|\mathbf{W}_{\otimes L}} &= \frac{P_{\mathbf{W}_{\otimes L}|X_i,Y_i}\,P_{X_i,Y_i}}{P_{\mathbf{W}_{\otimes L}}} = \frac{\mathcal{N}_{\mathbf{W}_{\otimes L}}(\mu_{\mathbf{W}_{\otimes L}|i}, \mathbf{\Sigma}_{\mathbf{W}_{\otimes L}|i})}{\mathcal{N}_{\mathbf{W}_{\otimes L}}(\mu_0, \frac{\sigma_0^2}{n}\mathbf{I}_{d_0})} \times \frac{1}{2}\mathcal{N}_{X_i}(Y_i\mu_0, \sigma_0^2\mathbf{I}_{d_0}) \\
&= \frac{1}{2}\mathcal{N}_{X_i}(Y_i\mu_0,\sigma_0^2\mathbf{I}_{d_0}) \times C_i\mathcal{N}_{\mathbf{W}_{\otimes L}}\left(Y_iX_i, \frac{(n-1)\sigma_0^2}{n}\mathbf{I}_{d_0}\right) = \frac{1}{2}\mathcal{N}_{Y_iX_i}\left(\mathbf{W}_{\otimes L}^\intercal, \frac{(n-1)\sigma_0^2}{n}\mathbf{I}_{d_0}\right),
\end{aligned}$$

where $C_i = n^{d_0}\sqrt{(\frac{2\pi\sigma^2}{n^2})^{d_0}}\exp\{\frac{1}{2\sigma_0^2}(Y_iX_i-\mu_0)^\intercal(Y_iX_i-\mu_0)\}$. By integrating $P_{X_i,Y_i|\mathbf{W}_{\otimes L}}$ over X_i, we obtain $P_{Y_i|\mathbf{W}_{\otimes L}} = \frac{1}{2}$. We can also conclude that $P_{X_i,Y_i|\mathbf{W},\mathbf{W}_{\otimes L}} = P_{X_i,Y_i|\mathbf{W}_{\otimes L}}$ and $P_{Y_i|\mathbf{W},\mathbf{W}_{\otimes L}} = P_{Y_i|\mathbf{W}_{\otimes L}} = \frac{1}{2}$.

Next, we compute the divergences between the prior and posterior.

Proof of Proposition 4. For any $l\in[L]$, conditioned on $(Y_i,\mathbf{W},\mathbf{W}_{\otimes L})$, the distribution of $T_{l,i} = \mathbf{W}_{\otimes l}X_i$ is Gaussian with the mean and covariance

$$\mathbb{E}[T_{l,i}|Y_i,\mathbf{W},\mathbf{W}_{\otimes L}] = Y_i\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes L}^\intercal, \quad \mathsf{Cov}[T_{l,i}|Y_i,\mathbf{W},\mathbf{W}_{\otimes L}] = \frac{(n-1)\sigma_0^2}{n}\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^\intercal.$$

Similarly for a test data sample, when conditioned on $(Y,\mathbf{W},\mathbf{W}_{\otimes L})$, the distribution of $T_l = \mathbf{W}_{\otimes l}X$ is Gaussian with mean and covariance

$$\mathbb{E}[T_l|Y,\mathbf{W},\mathbf{W}_{\otimes L}] = \mathbb{E}[T_l|Y,\mathbf{W}] = Y_i\mathbf{W}_{\otimes l}\mu_0, \quad \mathsf{Cov}[T_l|Y_l,\mathbf{W},\mathbf{W}_{\otimes L}] = \sigma_0^2\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^\intercal.$$

Note that when $\mathbf{W}_{\otimes l}$ is not a full-rank matrix, the covariance matrices $\mathsf{Cov}[T_{l,i}|Y_i,\mathbf{W},\mathbf{W}_{\otimes L}]$ and $\mathsf{Cov}[T_l|Y_l,\mathbf{W},\mathbf{W}_{\otimes L}]$ are singular. Thus, the posteriors $P_{T_{l,i}|Y_i,\mathbf{W},\mathbf{W}_{\otimes L}}$ and $P_{T_l|Y,\mathbf{W},\mathbf{W}_{\otimes L}}$ are the push-forwards of two d_l-dimensional nonsingular Gaussian distributions to the lower-dimensional space. In fact, the dimension can be further proven to be $\mathrm{rank}(\mathbf{W}_{\otimes l})$ via eigendecomposition. Take the test data sample $(X, Y=1)$ for example in the followings. Let $X = \mu_0 + Z$ and the l^{th} representation $T_l = \mathbf{W}_{\otimes l}X = \mathbf{W}_{\otimes l}\mu_0 + \mathbf{W}_{\otimes l}Z$, where $Z\sim\mathcal{N}(0,\sigma_0^2\mathbf{I}_{d_0})$ and $\mathbf{W}_{\otimes l}Z \sim \mathcal{N}(0,\sigma_0^2\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^\intercal)$ is a singular Gaussian distribution. Since $\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^\intercal$ is a $(d_l\times d_l)$ positive semi-definite and symmetric matrix, the eigendecomposition is given by $\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^\intercal = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\intercal$, where $\mathbf{U}$ is a $(d_l\times d_l)$ orthogonal matrix, $\mathbf{\Lambda}$ is a $(d_l\times d_l)$ diagonal matrix whose entries are the eigenvalues of $\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^\intercal$. We assume that $\mathbf{W}_{\otimes l}$ satisfies that only the first $\mathrm{rank}(\mathbf{W}_{\otimes l})$ entries of Λ are non-zero. Construct the following two d_l-dimenstional vectors:

$$Z_0 \coloneqq (Z, \underbrace{0,\dots,0}_{(d_l-d_0)\text{ entries}})^\intercal, \quad \tilde{Z} \coloneqq (\underbrace{\tilde{Z}_1,\dots,\tilde{Z}_{\mathrm{rank}(\mathbf{W}_{\otimes l})}}_{\sim\mathcal{N}(0,\sigma_0^2\mathbf{I}_{\mathrm{rank}(\mathbf{W}_{\otimes l})})}, \underbrace{0,\dots,0}_{(d_l-\mathrm{rank}(\mathbf{W}_{\otimes l}))\text{ entries}}).$$

Since the last $(d_l - \mathrm{rank}(\mathbf{W}_{\otimes l}))$ columns of $\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}$ are all zero, the following equalities hold:

$$\mathbf{W}_{\otimes l}Z \overset{\mathrm{d}}{=\!=} \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}Z_0 \overset{\mathrm{d}}{=\!=} \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\tilde{Z}, \quad \text{and} \quad T_l \overset{\mathrm{d}}{=\!=} \mathbf{W}_{\otimes l}\mu_0 + \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\tilde{Z}.$$

Thus, only the a subset of $\mathrm{rank}(\mathbf{W}_{\otimes l})$) covariates of T_l are effective random variables and the Gaussian PDF of T_l is on the $\mathrm{rank}(\mathbf{W}_{\otimes l})$-dimensional space. Let $r_l = \mathrm{rank}(\mathbf{W}_{\otimes l})$. We can define a restriction of Lebesgue measure to the $\mathrm{rank}(\mathbf{W}_{\otimes l})$-dimensional affine subspace of $\mathbb{R}^{r_l}$ where the Gaussian distribution is supported. With respect to this measure the distribution of T_l given $(\mathbf{W},\mathbf{W}_{\otimes l}, Y=1)$ has the density of the following motif:

$$p_{\mathbf{W},\mathbf{W}_{\otimes l},Y=1}(T_l) = \frac{\exp\left(-\frac{1}{2\sigma_0^2}(T_l-\mathbf{W}_{\otimes l}\mu_0)^\intercal(\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^\intercal)^\dagger(T_l-\mathbf{W}_{\otimes l}\mu_0)\right)}{\sqrt{(2\pi)^{r_l}\det^*(\sigma_0^2\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^\intercal)}} \tag{4}$$

where $(\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^\intercal)^\dagger = \mathbf{U}\mathbf{\Lambda}^\dagger\mathbf{U}^\intercal$ is the generalized inverse of $\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^\intercal$,$\mathbf{\Lambda}^\dagger$ is the pseudo-inverse of $\mathbf{\Lambda}$, and $\det^*$ is the pseudo-determinant. In a similar manner, the density of $T_{l,i}$ can obtained by replacing the mean $\mathbf{W}_{\otimes l}\mu_0$ with $\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes L}^\intercal$ and σ_0^2 with $\frac{(n-1)\sigma_0^2}{n}$.

Recall that the KL divergence between any two Gaussian distributions $P = \mathcal{N}(\mu_p, \mathbf{\Sigma}_p), Q = \mathcal{N}(\mu_q, \mathbf{\Sigma}_q) \in \mathcal{P}(\mathbb{R}^k)$ for some $k \in \mathbb{N}$ is given by

$$\mathsf{D_{KL}}(P\|Q) = \frac{1}{2}\left(\log\frac{\det^*(\mathbf{\Sigma}_q)}{\det^*(\mathbf{\Sigma}_p)} - k + (\mu_p - \mu_q)^\intercal \mathbf{\Sigma}_q^\dagger(\mu_p - \mu_q) + \operatorname{tr}(\mathbf{\Sigma}_q^\dagger \mathbf{\Sigma}_p)\right).$$

For $l = 1, \ldots, L$, the $\mathsf{UB}(l)$ in Theorem 1 can be written as

$$\mathsf{UB}(l) = \frac{\sqrt{2}\sigma}{n}\sum_{i=1}^n \sqrt{\mathsf{D_{KL}}(P_{T_{l,i},Y_i|\mathbf{W}}\|P_{T_l,Y|\mathbf{W}}|P_{\mathbf{W}})} \le \frac{\sqrt{2}\sigma}{n}\sum_{i=1}^n \sqrt{\mathsf{D_{KL}}(P_{T_{l,i},Y_i|\mathbf{W},\mathbf{W}_{\otimes L}}\|P_{T_l,Y|\mathbf{W},\mathbf{W}_{\otimes L}}|P_{\mathbf{W},\mathbf{W}_{\otimes L}})} =: \widetilde{\mathsf{UB}}(l),$$

where the hierarchical structure $\widetilde{\mathsf{UB}}(L) \le \widetilde{\mathsf{UB}}(L-1) \le \cdots \le \widetilde{\mathsf{UB}}(0)$ still holds.

From the probability density function (PDF) of T_l and $T_{l,i}$ (c.f. (4)), the KL divergence term in the upper bound $\widetilde{\mathsf{UB}}(l)$ can be rewritten as: for $l = 1, \ldots, L$,

$$\begin{aligned}
&\mathsf{D_{KL}}(P_{T_{l,i},Y_i|\mathbf{W},\mathbf{W}_{\otimes L}}\|P_{T_l,Y|\mathbf{W},\mathbf{W}_{\otimes L}}|P_{\mathbf{W},\mathbf{W}_{\otimes L}}) \\
&= \frac{1}{2}\mathsf{D_{KL}}(P_{T_{l,i}|Y_i=1,\mathbf{W},\mathbf{W}_{\otimes L}}\|P_{T_l|Y=1,\mathbf{W},\mathbf{W}_{\otimes L}}|P_{\mathbf{W},\mathbf{W}_{\otimes L}}) + \frac{1}{2}\mathsf{D_{KL}}(P_{T_{l,i}|Y_i=-1,\mathbf{W},\mathbf{W}_{\otimes L}}\|P_{T_l|Y=-1,\mathbf{W},\mathbf{W}_{\otimes L}}|P_{\mathbf{W},\mathbf{W}_{\otimes L}}) \\
&= \frac{1}{2}\mathbb{E}\left[\mathbb{E}\left[r_l\left(\log\frac{n}{n-1} - 1 + \frac{n-1}{n}\right) + \frac{1}{\sigma_0^2}(\mu_0 - \mathbf{W}_{\otimes L}^\intercal)^\intercal(\mathbf{W}_{\otimes l})^\intercal(\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^\intercal)^\dagger \mathbf{W}_{\otimes l}(\mu_0 - \mathbf{W}_{\otimes L}^\intercal)\middle|\mathbf{W},\mathbf{W}_{\otimes L}\right]\right] \\
&= \frac{1}{2}\mathbb{E}\left[\mathbb{E}\left[r_l\left(\log\frac{n}{n-1} - \frac{1}{n}\right) + \frac{1}{\sigma_0^2}(\mu_0 - \mathbf{W}_{\otimes L}^\intercal)^\intercal \mathbf{V}\mathbf{S}^\intercal\mathbf{U}^\intercal\mathbf{U}\mathbf{\Lambda}^\dagger\mathbf{U}^\intercal\mathbf{U}\mathbf{S}\mathbf{V}^\intercal(\mu_0 - \mathbf{W}_{\otimes L}^\intercal)\middle|\mathbf{W},\mathbf{W}_{\otimes L}\right]\right] \\
&= \mathbb{E}\left[\frac{r_l}{2}\left(\log\frac{n}{n-1} - \frac{1}{n}\right) + \frac{1}{2\sigma_0^2}\mathbb{E}\left[\|\mu_0 - \mathbf{W}_{\otimes L}^\intercal\|^2\right]\right] \\
&= \frac{\mathbb{E}[r_l]}{2}\left(\log\frac{n}{n-1} - \frac{1}{n}\right) + \frac{d_0}{2n},
\end{aligned}$$

where the last equality follows since $\frac{\sqrt{n}}{\sigma_0}(\mu_0 - \mathbf{W}_{\otimes L}^\intercal) \sim \mathcal{N}(0, \mathbf{I}_{d_0})$, $\frac{n}{\sigma_0^2}\|\mu_0 - \mathbf{W}_{\otimes L}^\intercal\|^2 \sim \chi_{d_0}^2$ and $\frac{1}{2\sigma_0^2}\mathbb{E}\left[\|\mu_0 - \mathbf{W}_{\otimes L}^\intercal\|^2\right] = \frac{d_0}{2n}$.

The KL divergence term in the upper bound $\widetilde{\mathsf{UB}}(0)$ can be rewritten as

$$\begin{aligned}
\mathsf{D_{KL}}(P_{X_i,Y_i|\mathbf{W},\mathbf{W}_{\otimes L}}\|P_{X,Y}|P_{\mathbf{W},\mathbf{W}_{\otimes L}}) &= \mathbb{E}\left[\mathbb{E}\left[\mathsf{D_{KL}}\left(\frac{1}{2}\mathcal{N}_{X_i}\left(-\mathbf{W}_{\otimes L}^\intercal, \frac{(n-1)\sigma_0^2}{n}\mathbf{I}_{d_0}\right)\middle\|\frac{1}{2}\mathcal{N}_X(-\mu_0, \sigma_0^2\mathbf{I}_{d_0})\right)\right.\right. \\
&\qquad \left.\left. + \mathsf{D_{KL}}\left(\frac{1}{2}\mathcal{N}_{X_i}\left(\mathbf{W}_{\otimes L}^\intercal, \frac{(n-1)\sigma_0^2}{n}\mathbf{I}_{d_0}\right)\middle\|\frac{1}{2}\mathcal{N}_X(\mu_0, \sigma_0^2\mathbf{I}_{d_0})\right)\middle|\mathbf{W},\mathbf{W}_{\otimes L}\right]\right] \\
&= \frac{1}{2}\mathbb{E}\left[\mathbb{E}\left[d_0\left(\log\frac{n}{n-1} - 1 + \frac{n-1}{n}\right) + \frac{1}{\sigma_0^2}\|\mu_0 - \mathbf{W}_{\otimes L}^\intercal\|^2\middle|\mathbf{W},\mathbf{W}_{\otimes L}\right]\right] \\
&= \frac{d_0}{2}\left(\log\frac{n}{n-1} - \frac{1}{n}\right) + \frac{1}{2\sigma_0^2}\mathbb{E}\left[\|\mu_0 - \mathbf{W}_{\otimes L}^\intercal\|^2\right] \\
&= \frac{d_0}{2}\left(\log\frac{n}{n-1} - \frac{1}{n}\right) + \frac{d_0}{2n}.
\end{aligned}$$

□

Proof of Proposition 5. Since the closed form of W_1 between two Gaussian distributions is not known but known for W_2 and $\mathsf{W}_1(\cdot,\cdot) \le \mathsf{W}_2(\cdot,\cdot)$, we consider analysing $\mathsf{W}_2(P_{T_{l,i},Y_i|\mathbf{W}}, P_{T_l,Y|\mathbf{W}}|P_{\mathbf{W}})$ as a surrogate of the upper bound in Theorem 2. Following the proof of Theorem 2, we can obtain

$$\begin{aligned}
\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y}) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\ell(\mathbf{W},X,Y) - \ell(\mathbf{W},X_i,Y_i)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbb{E}[\ell(\mathbf{W},X,Y) - \ell(\mathbf{W},X_i,Y_i)]|\mathbf{W},\mathbf{W}_{\otimes L},Y_i] \\
&\overset{(a)}{\le} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\bar{\rho}_0(\mathbf{W})\mathsf{W}_1(P_{X_i|Y_i,\mathbf{W},\mathbf{W}_{\otimes L}}, P_{X|Y,\mathbf{W},\mathbf{W}_{\otimes L}})\right] \\
&\le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\bar{\rho}_0(\mathbf{W})\mathsf{W}_2(P_{X_i|Y_i,\mathbf{W},\mathbf{W}_{\otimes L}}, P_{X|Y,\mathbf{W},\mathbf{W}_{\otimes L}})\right],
\end{aligned}$$

where (a) follows since $P_{Y_i|\mathbf{W},\mathbf{W}_{\otimes L}} = P_Y = \mathrm{Unif}\{-1,+1\}$. Similarly, we also have for all $l = 1,\ldots,L$.

$$\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y}) \le \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\bar{\rho}_l(\mathbf{W})\mathsf{W}_2(P_{T_{l,i}|Y_i,\mathbf{W},\mathbf{W}_{\otimes L}}, P_{T_l|Y,\mathbf{W},\mathbf{W}_{\otimes L}})\big].$$

By plugging the $P_{T_{l,i}|Y_i,\mathbf{W},\mathbf{W}_{\otimes L}}$ and $P_{T_l|Y,\mathbf{W},\mathbf{W}_{\otimes L}}$ (c.f. (4)) into the upper bound, we have

$$\begin{aligned}\mathsf{W}_2(P_{T_{l,i}|Y_i,\mathbf{W},\mathbf{W}_{\otimes L}}, P_{T_l|Y,\mathbf{W},\mathbf{W}_{\otimes L}}) &= \Bigg(\|\mathbf{W}_{\otimes l}(\mathbf{W}_{\otimes L}^{\intercal}-\mu_0)\|^2 + \mathrm{tr}\Big(\frac{(n-1)\sigma_0^2}{n}(\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^{\intercal}) + \sigma_0^2(\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^{\intercal}) \\ &\qquad - 2\Big(\frac{(n-1)\sigma_0^4}{n}(\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^{\intercal})^{\frac12}(\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^{\intercal})(\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^{\intercal})^{\frac12}\Big)^{\frac12}\Big)\Bigg)^{\frac12} \\ &= \sqrt{\|\mathbf{W}_{\otimes l}(\mathbf{W}_{\otimes L}^{\intercal}-\mu_0)\|^2 + \frac{(\sqrt{n-1}-\sqrt{n})^2\sigma_0^2}{n}\|\mathbf{W}_{\otimes l}\|_{\mathrm{F}}^2}.\end{aligned}$$

Similarly, we have

$$\mathsf{W}_2(P_{X_i|Y_i,\mathbf{W},\mathbf{W}_{\otimes L}}, P_{X|Y,\mathbf{W},\mathbf{W}_{\otimes L}}) = \sqrt{\|(\mathbf{W}_{\otimes L}^{\intercal}-\mu_0)\|^2 + \frac{d_0(\sqrt{n-1}-\sqrt{n})^2\sigma_0^2}{n}}.$$

Since the activation functions are all 1-Lipschitz, i.e., $\rho_l = 1$ for all $l = 1,\ldots,L$ and the loss function $\tilde{\ell}$ is $4\sqrt{2}$-Lipschitz, we have $\bar{\rho}_l(\mathbf{W}) = 4\sqrt{2}\big(1 \vee \prod_{j=l+1}^{L}\|\mathbf{W}_j\|_{\mathsf{op}}\big)$ for $l = 0,1,\ldots,L$.

For notational simplicity, let $\mathbf{W}_{\otimes 0} = \mathbf{W}_0 = \mathbf{I}_{d_0}$ and $r_0 = d_0$. Here we use $\|\cdot\|_{\mathrm{F}}$ of a vector to equivalently denote its Euclidean norm, with a slight abuse of notations. Then the generalization error is upper bounded by

$$\begin{aligned}&\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y}) \\ &\le \min\Bigg\{\min_{l=1,\ldots,L}\mathbb{E}\Bigg[\bar{\rho}_l(\mathbf{W})\sqrt{\|\mathbf{W}_{\otimes l}(\mathbf{W}_{\otimes L}^{\intercal}-\mu_0)\|^2 + \frac{(\sqrt{n-1}-\sqrt{n})^2\sigma_0^2}{n}\|\mathbf{W}_{\otimes l}\|_{\mathrm{F}}^2}\Bigg], \\ &\qquad \mathbb{E}\Bigg[\bar{\rho}_0(\mathbf{W})\sqrt{\|(\mathbf{W}_{\otimes L}^{\intercal}-\mu_0)\|^2 + \frac{d_0(\sqrt{n-1}-\sqrt{n})^2\sigma_0^2}{n}}\Bigg]\Bigg\} \\ &\overset{(a)}{\le} \min\Bigg\{\min_{l=1,\ldots,L}\mathbb{E}\Bigg[\bar{\rho}_l(\mathbf{W})\Bigg(\|\mathbf{W}_{\otimes l}(\mathbf{W}_{\otimes L}^{\intercal}-\mu_0)\| + \frac{(\sqrt{n}-\sqrt{n-1})\sigma_0}{\sqrt{n}}\|\mathbf{W}_{\otimes l}\|_{\mathrm{F}}\Bigg)\Bigg], \\ &\qquad \mathbb{E}\Bigg[\bar{\rho}_0(\mathbf{W})\Bigg(\|(\mathbf{W}_{\otimes L}^{\intercal}-\mu_0)\| + \frac{\sqrt{d_0}(\sqrt{n}-\sqrt{n-1})\sigma_0}{\sqrt{n}}\Bigg)\Bigg]\Bigg\} \\ &\overset{(b)}{\le} \min\Bigg\{\min_{l=1,\ldots,L}\mathbb{E}\Bigg[\bar{\rho}_l(\mathbf{W})\|\mathbf{W}_{\otimes l}\|_{\mathrm{F}}\Bigg(\|(\mathbf{W}_{\otimes L}^{\intercal}-\mu_0)\| + \frac{(\sqrt{n}-\sqrt{n-1})\sigma_0}{\sqrt{n}}\Bigg)\Bigg], \\ &\qquad \mathbb{E}\Bigg[\bar{\rho}_0(\mathbf{W})\sqrt{d_0}\Bigg(\|(\mathbf{W}_{\otimes L}^{\intercal}-\mu_0)\| + \frac{(\sqrt{n}-\sqrt{n-1})\sigma_0}{\sqrt{n}}\Bigg)\Bigg]\Bigg\} \\ &= \min_{l=0,\ldots,L}\mathbb{E}\Bigg[\bar{\rho}_l(\mathbf{W})\|\mathbf{W}_{\otimes l}\|_{\mathrm{F}}\Bigg(\|\mathbf{W}_{\otimes L}^{\intercal}-\mu_0\| + \frac{(\sqrt{n}-\sqrt{n-1})\sigma_0}{\sqrt{n}}\Bigg)\Bigg] \\ &\overset{(c)}{\le} \min_{l=0,\ldots,L}\mathbb{E}\left[\bar{\rho}_l(\mathbf{W})^2\|\mathbf{W}_{\otimes l}\|_{\mathrm{F}}^2\right]^{\frac12}\mathbb{E}\Bigg[\Bigg(\|\mathbf{W}_{\otimes L}^{\intercal}-\mu_0\| + \frac{(\sqrt{n}-\sqrt{n-1})\sigma_0}{\sqrt{n}}\Bigg)^2\Bigg]^{\frac12} \\ &\overset{(d)}{\le} \min_{l=0,\ldots,L}\mathbb{E}\left[\bar{\rho}_l(\mathbf{W})^2\|\mathbf{W}_{\otimes l}\|_{\mathrm{F}}^2\right]^{\frac12}\Bigg(\frac{\sqrt{d_0}\sigma_0}{\sqrt{n}} + \frac{(\sqrt{n}-\sqrt{n-1})\sigma_0}{\sqrt{n}}\Bigg) \\ &= \frac{4\sqrt{2}\sigma_0(\sqrt{d_0}+(\sqrt{n}-\sqrt{n-1}))}{\sqrt{n}}\min_{l=0,\ldots,L}\mathbb{E}\Bigg[\Bigg(1 \vee \prod_{j=l+1}^{L}\|\mathbf{W}_j\|\Bigg)^2\|\mathbf{W}_{\otimes l}\|_{\mathrm{F}}^2\Bigg]^{\frac12},\end{aligned}$$

where (a) follows since $\sqrt{a^2+b^2} \le |a|+|b|$, (b) follows from the Cauchy-Schwarz inequality, (c) follows from the Hölder's inequality, and (d) follows from Minkowski's inequality and $\frac{n}{\sigma_0^2}\|\mathbf{W}_{\otimes L}^{\intercal}-\mu_0\|^2 \sim \chi_{d_0}^2$. □

Proof of Example 1. Since $\mathbf{W}_l$ is (2×2) rotation matrix multiplied by a scalar factor C_l and $W_L = (0, C_L)$ is a row vector, $\|\mathbf{W}_l\| = C_l$ for $l = 1,\ldots,L$. We have $\bar{\rho}_l(\mathbf{W}) = 4\sqrt{2}\big(1 \vee \prod_{j=l+1}^{L}\|\mathbf{W}_j\|\big) = 4\sqrt{2}\big(1 \vee \prod_{j=l+1}^{L}C_j\big)$ for $l = 0,1,\ldots,L$.
□